约束调控结点的基因网络构建算法*

刘 飞
(宝鸡文理学院物理与光电技术学院 宝鸡 721016)

摘 要：从实验数据构建基因调控网络是计算生物学领域的一个研究热点，但是启发式搜索、基因最大父结点数量限制策略和条件最优搜索等构建方法的计算复杂度都较大。启发式搜索方法的缺陷众所周知，在启发式搜索策略中很少有人限制基因结点的父结点数量，且搜索结果为了达到最优使得算法的时间复杂度变得很高。通过理论分析和实验结果，说明了最大父结点数量选取问题的优点和缺点，然后利用最优搜索方法融合最大父结点数量选取优点和已知基因调控网络拓扑信息，提出了新的基因调控网络构建方法。该方法利用贝叶斯网络框架实现，并在不同规模和拓扑结构的生物分子数据，真实网络数据和计算机人工合成数据集上进行测试，实验结果显示，该方法比现存的最优搜索算法有更快的计算速度。

关键词：贝叶斯网络；网络构建；基因调控网络；计算复杂度
中图分类号：TN01 **文献标识码**：A **国家标准学科分类代码**：520.60

Inferring gene networks with constrained regulatory nodes

Liu Fei
(Institute of Physics and Optoelectronics Technology, Baoji University of Arts and Science, Baoji 721016, China)

Abstract: Inferring the gene regulatory network (GRN) structure from data is an important problem in computational biology. However, it is a computationally complex problem and approximate methods such as heuristic search techniques, restriction of the maximum-number-of-parents (maxP) for a gene, or an optimal search under special conditions are required. The limitations of a heuristic search are well known but literature on the detailed analysis of the widely used maxP technique is lacking. The optimal search methods require large computational time. We report the theoretical analysis and experimental results of the strengths and limitations of the maxP technique. Further, using an optimal search method, we combine the strengths of the maxP technique and the known GRN topology to propose a novel algorithm. This algorithm is implemented in a Bayesian network framework and tested on biological, realistic, and in silico networks of different sizes and topologies. It can overcome the limitations of the maxP technique and show superior computational speed when compared to the current optimal search algorithms.

Keywords: Bayesian network; network construction; gene regulatory networks; computational complexity

1 引 言

理解复杂的基因调控网络对现代生物医学研究有很重要的作用。随着高通量 DNA 微阵列数据的产生，诞生了大量的 GRN 构建方法和模型[1-3]。基于共表达模型算法的优点是过程简单，计算复杂度低，适合大规模网络，但是其缺点是推断的相互作用关系没有体现因果性和系统动态性。基于常微分方程模型能很好地表示系统动态特性，但是该模型非常依赖实验参数，且有很高的计算复杂度，只适用于小规模网络。贝叶斯网络(Bayesian network, BN)和动态贝叶斯网络是一种概率图模型，其构建网络的计算复杂度和网络规模上处于上述两种方法之间。BN 是基于概率和统计且非常流行的一种 GRN 构建方法，因为它可以从实验数据中学习网络结点间因果关系，并且对实验数据噪声有很大的鲁棒性，除此之外 DBN 还可以用实验时序数据给反馈环建模。

随着网络中基因结点数量的增加，从中构建网络模型已经成为一个复杂性计算问题，因为从所有可能的网络结构中搜索一个最优网络结构是一个 NP 难问题[4-6]。启发式搜索和基因结点的 maxP 技术是两个常用的迭代策略。启发式搜索方法的缺陷众所周知，该方法不能保证得到的结构是全局最优。对于 maxP 技术来说，一个基因结点的

收稿日期：2016-11
* 基金项目：宝鸡市科技计划(15RKX-1-5-18)、宝鸡文理学院科研(ZK16016，ZK16032)项目资助

父节点数量的取值是随意确定，当取值较小时，算法计算复杂度就较小。然而从 GRN 拓扑信息和基因结点组合调控先验知识可以得出某些个别基因结点可能被很多个基因结点所调控[7]，那么随意确定其父节点的数量就会限制预测网络中的边数，从而影响网络构建的精度。但是 maxP 技术不需要一些生物网络标签先验知识，作为一种通用的方法被广泛应用到各种 GRN 构建算法模型中，且很有效地降低了算法的计算复杂度。但是每个基因结点父节点数量的取值范围到底对 GRN 构建准确度有多大影响，本文详细地讨论了这个问题。

BNfinder[8]和 globalMIT[9]方法采用最优动态贝叶斯结构学习来构建 GRN，为了克服最优搜索方法计算时间复杂度高的缺陷，这些方法都采用分解打分函数，并且假定一些优先搜索，克服启发式搜索的不确定性[10-11]。但是这些方法都只适合小规模网络，因为要确定一个基因结点所有可能的组合调控是非常困难的。所以本文借助一些 GRN 的拓扑先验知识，如某些基因结点的父节点很少来降低计算时间复杂度，并且和第二种方法进行了比较，取得了较好的效果。

2　贝叶斯理论方法

对于一个随机变量 $X=\{X_1,X_1,\cdots,X_n\}$，贝叶斯网络就是这些变量之间概率统计关系的一个图模型，而且它是一个有向无环图 G(directed acyclic graph, DAG)。在贝叶斯网络中，顶点(结点)就是这些随机变量(基因)，边就是这些随机变量(基因)之间的概率依赖。假如在结点 A 到结点 B 有一条有向边，那么就称结点 A 是结点 B 的父亲，结点 B 是结点 A 的孩子。一个结点在给定父结点的情况下，根据 Markov 假设，这个结点和它的非子孙结点都是相互独立的，可以用 $P(X_1,X_1,\cdots,X_n)$ 这个联合概率分布来表示。基于图模型，他可以分解为一系列条件概率的乘积，表达式如下：

$$P(X_1,X_1,\cdots,X_n)=\prod_{X_i\in X}P(X_i\mid Pa(X_i)) \tag{1}$$

其中 $Pa(X_i)$ 为图 G 中结点 X_i 的父结点集合。图 1 是一个简单的贝叶斯网络实例，它揭示了癌症变量(C)、环境变量(E)、生物标志物(B)和 3 种单核苷酸多态性(SNPs)之间的关系。这几个变量的联合概率分布可以定义为如下表达式。

$$P(S1,S2,S3,E,B,C)=P(B\mid E)\cdot P(C\mid E,S1)\cdot P(E\mid S3)\cdot P(S1S2)\cdot P(S3)\cdot P(S2) \tag{2}$$

在贝叶斯网络结构学习中，在数据集 D 上，通过贝叶斯打分策略全局搜索找出最可能的图结构 G，也就是找出所有可能的图结构(找出基因结点之间所有可能的边)，其中贝叶斯打分最高的那个图结构作为基因调控网络的结构。可是随着结点(基因)个数的增加，其可能的网络拓扑结构会呈指数级增长，要想从所有可能的图结构中找出打分最优的图结构，这是一个典型的 NP-难问题。一个大规模的数据集 D 可能包括许多结点(基因)，给所有可能的图结构 G 都进行贝叶斯打分，它的时间复杂度是非常大的。因此，一般都采用 greedy-hill climbing 算法、Markov Chain Monte Carlo 方法和 simulated annealing 等启发式搜索算法来减小时间复杂度，并进行贝叶斯网络的结构学习[12]。

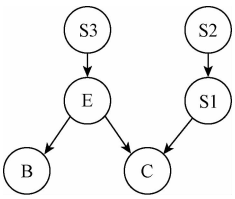

图 1　一个简单的贝叶斯网络图例

在本文中采用互信息测试(Mutual Information Test, MIT)来对贝叶斯网络进行打分[13]，设 $\boldsymbol{X}=\{X_1,X_2,\cdots,X_n\}$ 表示 n 个变量结点，它们对应的样本分别为 $\{r_1,r_2,\cdots,r_n\}$，在数据集 D 中总共有 N 个样本值，G 表示贝叶斯网络，$Pa_i=\{X_{i1},X_{i2},\cdots,X_{is_i}\}$ 表示结点 X_i 所有的父结点集合，其对应的样本为 $\{r_{i1},r_{i2},\cdots,r_{is_i}\}$，$s_i$ 为父结点的个数，那么 MIT 打分函数定义如下：

$$S_{MIT}(G:D)=\sum_{i=1,Pa_i\neq\varnothing}^{n}\{2N\cdot MI(X_i,Pa_i)-\sum_{j=1}^{s_i}\chi_{\alpha,l_i\sigma_i(j)}\} \tag{3}$$

这里 $MI(X_i,Pa_i)$ 表示结点 X_i 和其父结点的互信息值估计。$\chi_{\alpha,l_{ij}}$ 表示卡方分布在显著性水平为 α 时的值，$l_i\sigma_i(j)$ 表达式定义如下：

$$l_i\sigma_i(j)=\begin{cases}(r_i-1)(r_{i\sigma_i(j)}-1)\prod_{k=1}^{j-1}r_{i\sigma_i(k)}, & j=2,\cdots,s_i\\(r_i-1)(r_{i\sigma_i(j)}-1), & j=1\end{cases} \tag{4}$$

其中 $\sigma_i=\{\sigma_i(1),\sigma_i(2),\cdots,\sigma_i(s_i)\}$ 为父结点 $Pa_i=\{X_{i1},X_{i2},\cdots,X_{is_i}\}$ 索引 $\{1,2,\cdots,s_i\}$ 的一个随机置换。

3　基因网络的拓扑分析

在基因调控网络中结点入度值呈指数级下降，也就是说在 GRN 中大部分基因结点只被少数几个父结点调控，只有个别基因结点的父结点数较大。大肠杆菌[14]、分枝杆菌[15]和枯草杆菌[16]GRN 中的数据可以证明这个观点，它们网络中结点的平均入度值分别为 2,1 和 1，结点的最大入度值分别为 17、11 和 16。表 1 显示了 3 种网络结点入度分别为 1、2、3 时结点所占总结点个数的百分比，据这 3 个不完全网络统计，GRN 中 79%以上基因结点只被 3 个或 3 个以下父结点调控，所以 GRN 中结点入度值呈指数级下降已经成为了一条先验知识。

从实验基因表达数据构建基因调控网络最大的挑战就是其可能的图结构数量随着结点数呈指数级增长。在图论

表 1　小入度结点百分比统计表　　(%)

入度	1	2	3	累计
大肠杆菌	35	28	16	79
枯草杆菌	52	28	11	91
分枝杆菌	81	15	3	99

中基因就是网络图中的结点,一个基因的调控基因称为基因结点的父结点,基因和调控基因之间的关系称为结点和其父结点之间的有向边,这样 GRN 就成为了一个有向图。对于一个有向图来说,如果有 n 个结点,那么图可能的结构就有 2^{n^2} 个,这个数量是非常巨大的,计算所有可能图结构的复杂度会随着结点数呈指数倍增长。

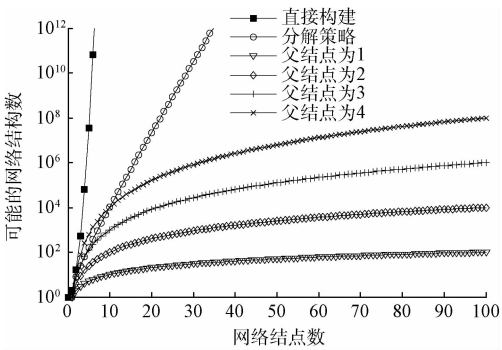

图 2　不同节点数可能的网络结构数

随着网络中结点个数的增加,其可能的网络结构数会呈指数级增长,如图 2 所示,总体分为 3 类:直接构建有向图;用分解打分函数策略构建有向图;用分解打分函数和限制父节点数量策略共同构建有向图。从图中可以看出对于相同的结点个数,可能的网络结构数从第 1 种方法到第 3 种方法迅速下降。面对中大规模网络,前两种方法的效率会非常低,但是第 3 种方法会随着网络结点数目增加其可能的网络结构数有一个明显地收敛。假设给一个网络打分会耗时 1 s,那么分析 3.15×10^7 个网络会用时 1 年,从图中可以看出前两种方法当节点数为 5 和 20 时其可能的网络结构数目近似达到 10^8,因此第 1 种方法当结点数大于 5 时几乎是不可行的,第 2 种方法在 20 个结点以内还是可行的。同样从图中可以得出 100 个结点网络(父节点数量为 2)的可能结构数为 10^4,分析这个网络大约需要一天时间,这大大降低了算法的复杂度。

使用分解打分策略的 BNFinder 和 globalMIT 算法,其性能良好且克服了过高的计算时间复杂度,这些方法应用了数理统计知识和原始假设,用卡方检验 α 计算结点间的统计独立性。虽然这种方法的性能较好,但是其可能的网络结构数依然很大,达到 $n2^n$ 种可能。采用这种传统打分分解策略的网络计算复杂度依然很大,采用 maxP 策略构建 GRN 时限定结点最大父结点数为 k,那么计算复杂度就变为 n^k,这个计算量相对较小。和上述方法计算时间复杂度的比值为 $n^k/n2^n = n^{k-1}2^{-n}$,这个比值就是计算时间复杂度的改进,比值越小改进的效果越好。其中 n 表示结点个数,k 表示最大父结点数值,可以看出当结点数为 10 时,父节点数为 1 时,采用 maxP 策略的方法比传统打分策略的方法在运算时间上快了 1 000 倍,随着结点数增长,算法的执行时间呈指数级下降。这些数据很好地阐述了采用 maxP 策略构建 GRN 时的优越性,但是这种方法存在一个缺陷,就是 k 值的选取不能太大,当 k 值太大时,算法的时间复杂度会急剧增长,然而部分结点的父结点调控数量肯定会大于 k,这势必会影响算法的效率和精度。

上述 maxP 策略在 k 值选取上的缺陷可以用一种迭代的方法来克服。在第一步迭代中,所有基因结点其父结点数量限定为给定的数值来计算;在第二步迭代中,找出那些有 k 个父结点的基因(认为这些结点会潜在的有更多结点调控),不要限制其父结点的数量重新计算其可能的调控的父结点,这样就会避免因限制 k 的数值使得一些结点父结点调控数量少而引起的假阴性,从而提高构建 GRN 算法的准确性,本文在构建基因调控网络就是采用这种方法。这种方法在第二步迭代中克服了 maxP 策略的缺陷,在第一步迭代中保留了 maxP 策略计算机复杂度小的优势,因为大部分基因结点只有少量的父结点调控,这些结点的计算只出现在第一步迭代中,从而减小了计算的复杂度,少数基因结点会有数量较多的父结点调控,而这些基因结点在第二步迭代重新计算,从而保证其结点调控数量的准确性。

4　实验仿真分析

为了说明本文算法的性能,在 IRMA 网络[17],SOS 网络,10 基因和 20 基因[18]网络数据上进行仿真验证。本文以 globalMIT 算法的时间计算复杂度为基准,当最大父结点数量为 1、2、3 和 4 时,本文算法分别在以上 4 个网络上进行计算,得出的相对运行时间值如图 3 所示。灰度柱状图从左至右依次是对 IRMA、SOS、10 基因和 20 基因网络数据在不同父结点的情况计算得出的相对运行时间。在父结点为 2 的情况下,这四种网络的平均运行时间为 0.29,比 globalMIT 算法时间复杂度快 3 倍多;在父结点为 1 的情况下,这四种网络的算法时间复杂度明显要高于父结点为 2 的情况,这是因为在本文算法在第 2 次迭代过程中结点的个数太多导致的;在父结点为 3 和 4 的情况下,这 4 种网络的算法时间复杂度明显要高于父结点为 1 和 2 的情况,这是因为本文算法第 1 次迭代的结点父结点个数太多导致了较高的时间复杂度,但是这个算法整体的时间复杂度相比 globalMIT 算法还是有较好的优越性。

本文算法在数据集 IRMA、SOS、10 基因和 20 基因网络上的性能相对 globalMIT 算法,在性能指标 AUC-PR 上的表现如图 4 所示。从图中可以看出,在 4 个不同的数据集上,当父结点个数为 1、2、3 和 4 时,本文算法的性能几乎和 globalMIT 算法性能一样,这说明了本文算法不仅在时

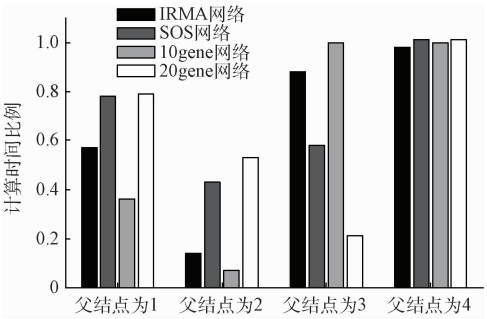

图 3　本文算法时间复杂度的相对比值

间性能上取得了较大的改进，而且在网络构建性能精度上也取得了较好的结果。

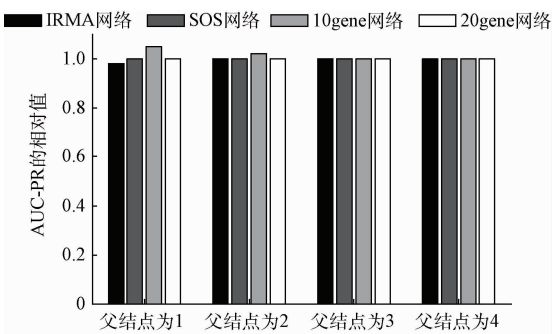

图 4　本文算法 AUC-PR 性能的相对比值

5　结　　论

用贝叶斯网络推断基因调控网络的最优结构是一个 NP 难问题。为了减少其计算时间复杂度，经常采用的一个策略就是限制每个基因结点的入度值(即限制每个基因结点的父结点数量问题，MaxP 技术)。本文从理论上和实际细节方面详细地分析了这个问题，实验结果表明实际的运行时间基本和理论计算时间相吻合。另外运行时间在父节点数量较小时取得了指数级的改善，时间复杂度也没有过分依赖网络的密集程度，只和网络的结点个数相关。

本文算法凭借 maxP 技术的优点减少了时间计算复杂度，凭借 GRN 拓扑结构先验知识减少每个节点的搜索空间，以此降低算法的时间复杂度。本文的方法和现在一些流行的 GRN 最优推断方法相比，可以取得相同的网络构建精度，但在时间复杂度上能降低很多。如在父结点数量为 2 时，其网络推断执行时间可以缩短为原来的 1/3。本文算法具有计算复杂度低，构建精度较高的优点，在中大规模网络推断中优势会更明显。

参考文献

[1] MADHAMSHETTIWAR P B, MAETSCHKE S R, DAVIS M J, et al. Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets[J]. Genome medicine, 2012, 4(5): 1-16.

[2] 刘春，马颖. 遗传算法和神经网络结合的 PSD 非线性校正[J]. 电子测量与仪器学报，2015，29(8)：1157-1163.

[3] LIU F, ZHANG S W, GUO W F, et al. Inference of gene regulatory network based on local bayesian networks[J]. PLoS Comput Biol, 2016, 12(8): e1005024.

[4] 李小珉，尹明. 基于遗传算法的 BP 神经网络电子系统状态预测方法研[J]. 电子测量技术，2016，39(9)：182-186.

[5] 李哲谦，刘书明，严壮志，等. 基于支持向量机的蛋白质相互作用预测[J]. 电子测量技术，2008，31(5)：4-8.

[6] 秦勇，梁旭. 基于混合遗传算法的并行测试任务调度研究[J]. 国外电子测量技术，2016，35(9)：72-75.

[7] ALBERT R. Scale-free networks in cell biology[J]. Journal of Cell Science, 2005, 118(21): 4947-4957.

[8] WILCZYŃSKI B, DOJER N. BNFinder: exact and efficient method for learning Bayesian networks[J]. Bioinformatics, 2009, 25(2): 286-287.

[9] VINH N X, CHETTY M, COPPEL R, et al. GlobalMIT: learning globally optimal dynamic bayesian network with the mutual information test criterion[J]. Bioinformatics, 2011, 27(19): 2765-2766.

[10] 张宇献，钱小毅，彭辉灯，等. 基于等位基因的实数编码量子进化算法[J]. 仪器仪表学报，2015，36(9)：2129-2137.

[11] XING H, LU C, ZHANG Q. Frequency modulated weak signal detection based on stochastic resonance and genetic algorithm[J]. Instrumentation, 2016, 3(1): 41-49.

[12] WU J, ZHAO X, LIN Z, et al. Large scale gene regulatory network inference with a multi-level strategy[J]. Molecular Biosystems, 2016, 12(2): 588-597.

[13] DE CAMPOS L M. A scoring function for learning Bayesian networks based on mutual information and conditional independence tests[J]. The Journal of Machine Learning Research, 2006(7): 2149-2187.

[14] SALGADO H, PERALTA-GIL M, GAMA-CASTRO S, et al. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more[J]. Nucleic Acids Research, 2013, 41(D1): 203-213.

(下转第 104 页)